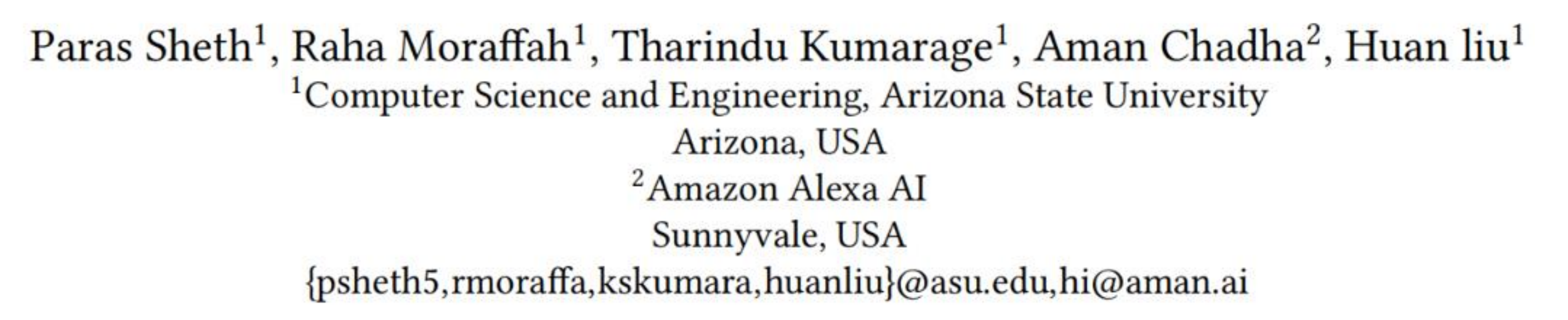
Causality Guided Disentanglement for Cross-Platform Hate Speech Detection



Paras Sheth[1], Raha Moraffah[1], Tharindu Kumarage[1], Aman Chadha[2], Huan liu[1]
[1]Computer Science and Engineering, Arizona State University
Arizona, USA
[2]Amazon Alexa AI
Sunnyvale, USA
{psheth5,rmoraffa,kskumara,huanliu}@asu.edu,hi@aman.ai

Code: https://github.com/paras2612/CATCH

—— WSDM 2024

2023.12.24 • ChongQing

Reported by Renhui Luo

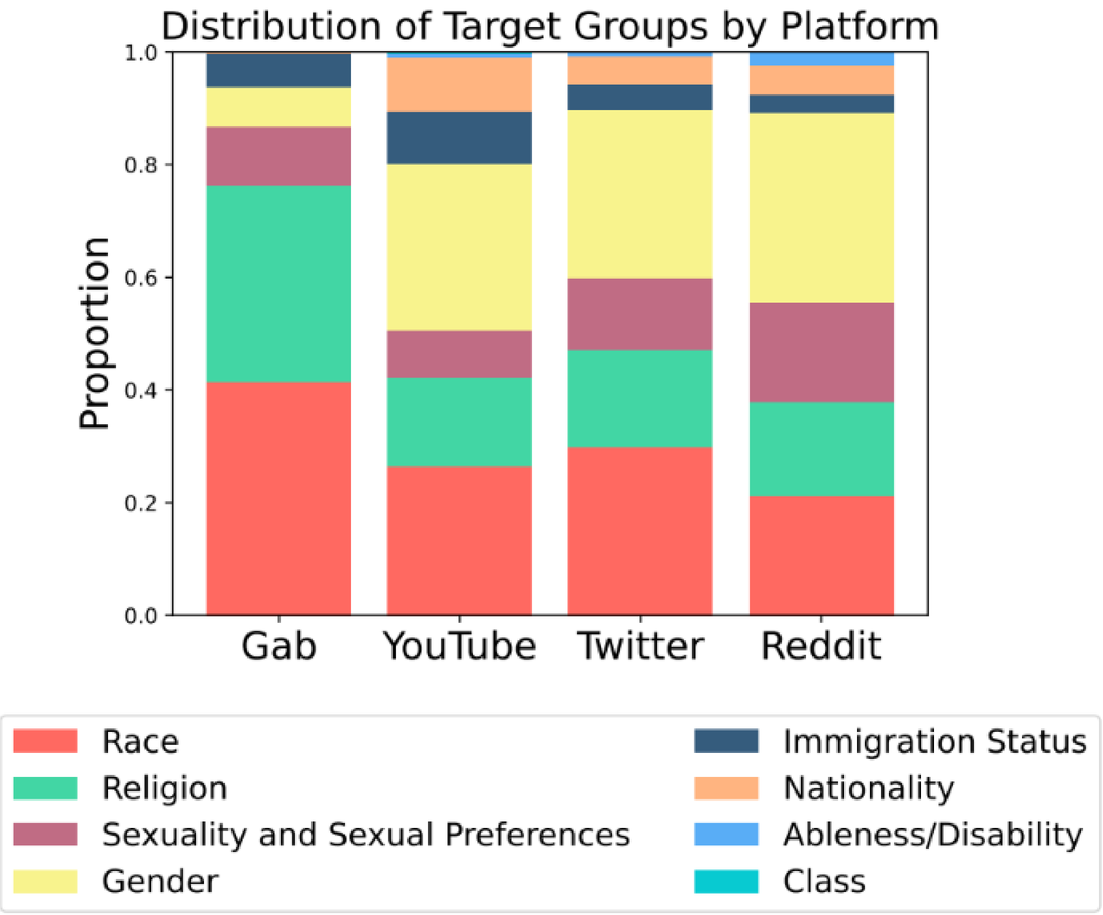
1.Introduction

2.Overview

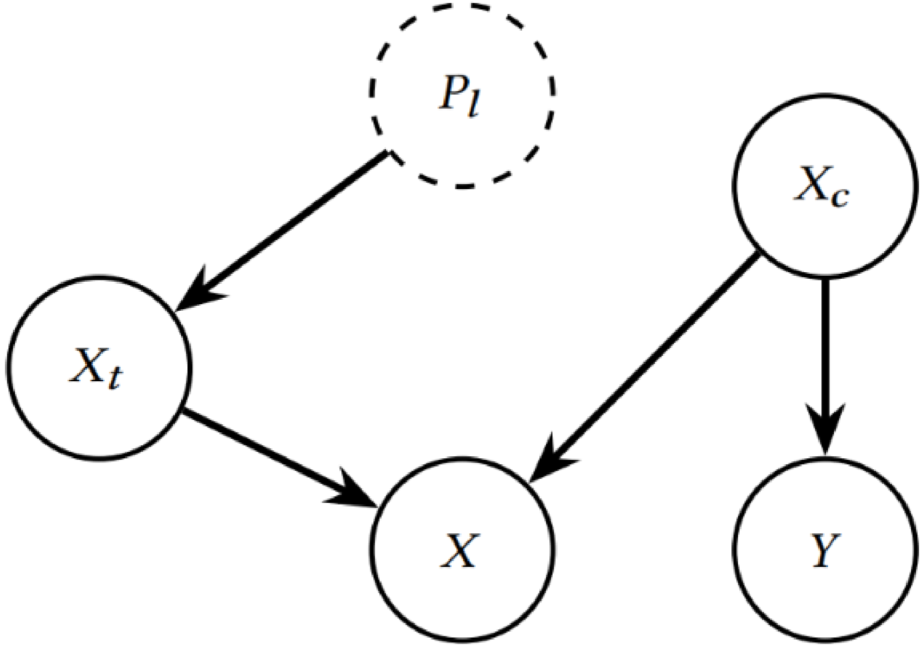
3.Methods

4.Experiments

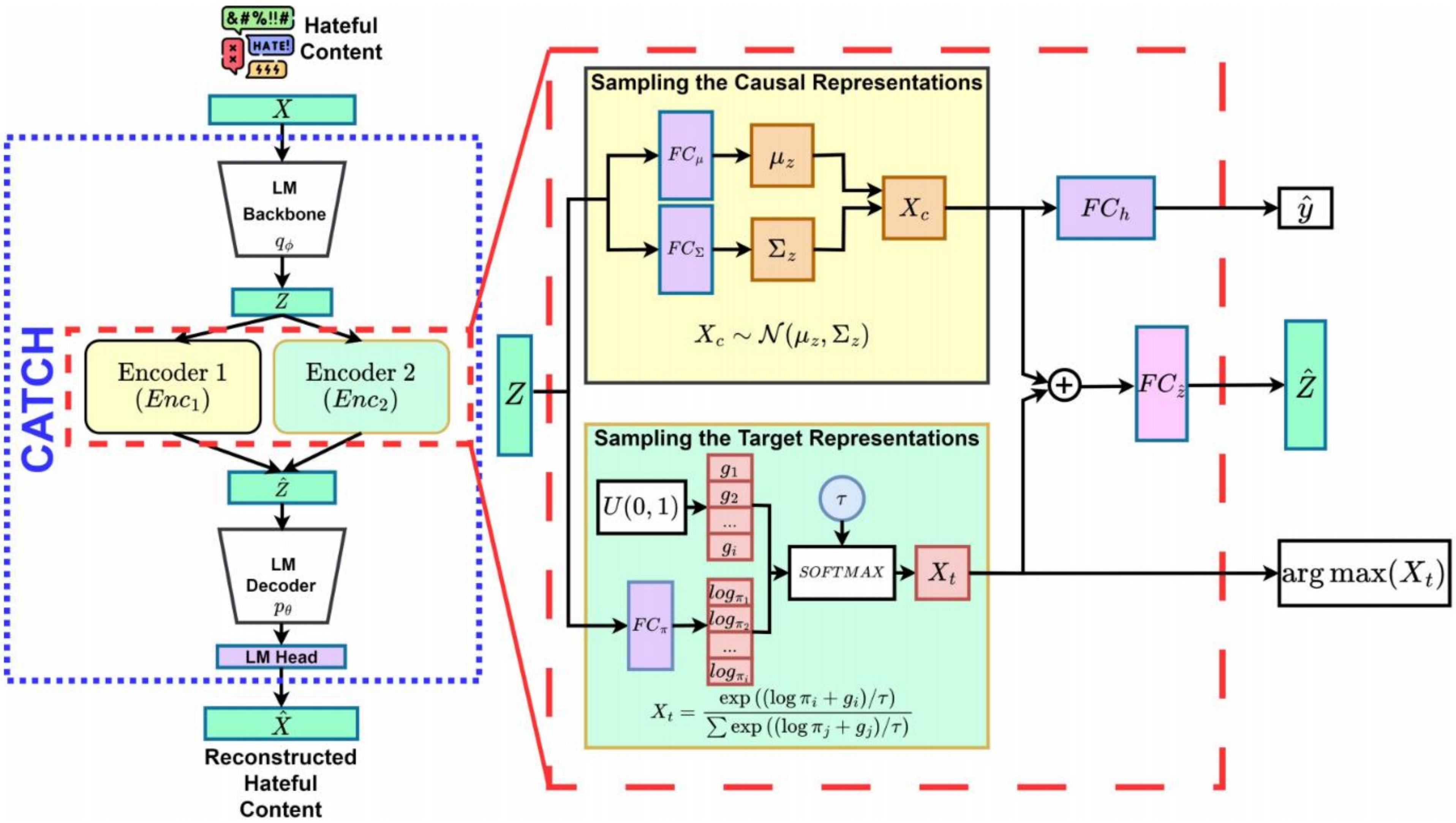
Introduction



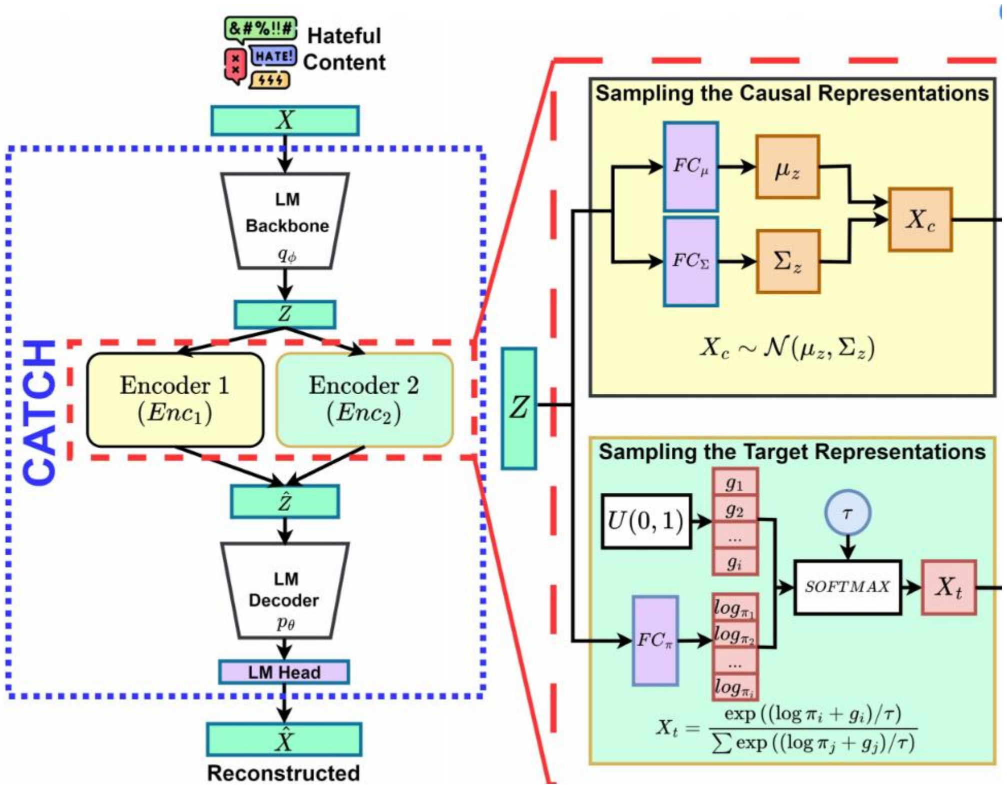
Introduction



Overview


&#%!!#
HATE!
Hateful Content
X
LM Backbone q_ϕ
Z
CATCH
Encoder 1 (Enc_1)
Encoder 2 (Enc_2)
$\hat{Z}$
LM Decoder p_θ
LM Head
$\hat{X}$
Reconstructed Hateful Content
Z
Sampling the Causal Representations
FC_μ
FC_Σ
μ_z
Σ_z
X_c
$X_c \sim \mathcal{N}(\mu_z, \Sigma_z)$
FC_h
$\hat{y}$
$FC_{\hat{z}}$
$\hat{Z}$
Sampling the Target Representations
$U(0,1)$
g_1
g_2
...
g_i
τ
SOFTMAX
X_t
FC_π
log_{π_1}
log_{π_2}
...
log_{π_i}
$X_t = \frac{\exp((\log \pi_i + g_i)/\tau)}{\sum \exp((\log \pi_j + g_j)/\tau)}$
$\arg\max(X_t)$

Method



$$z = q_\phi\left(\gamma(x)\right), \tag{1}$$

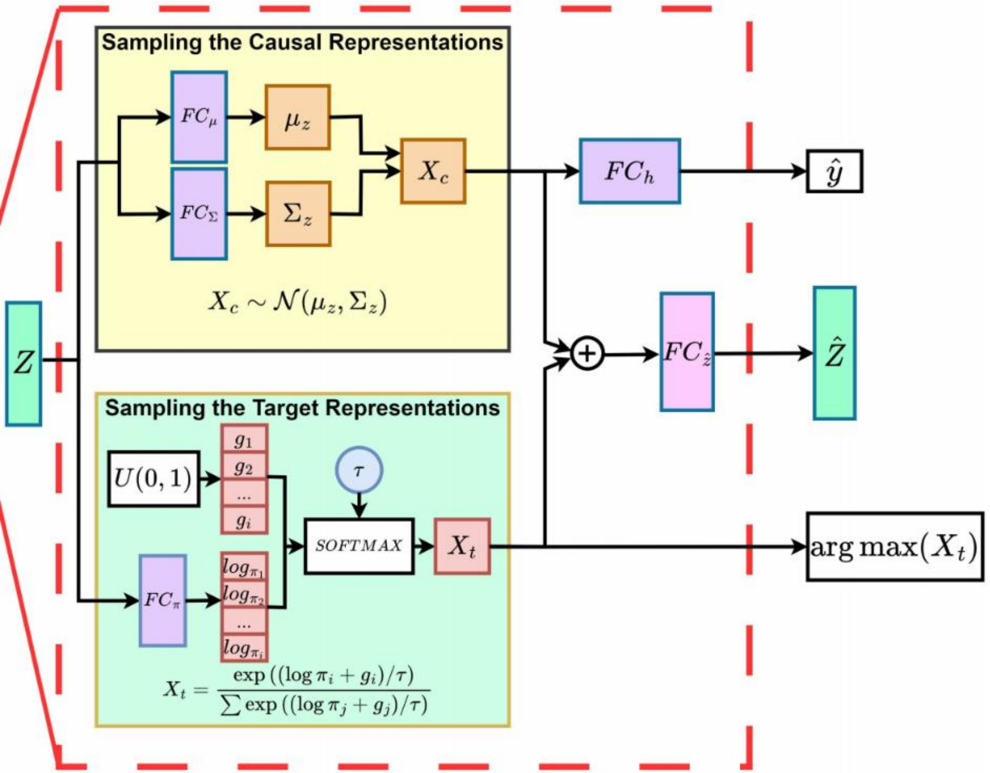
$$\mu_z = FC_\mu(z),\ \Sigma_z = FC_\Sigma(z),$$

$$X_c = Enc_1(\mu_z, \Sigma_z) = \mu_z + \Sigma_z \odot \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{2}$$

$$z_\pi = FC_\pi(z)$$

$$X_t = Enc_2(\pi, g) = \frac{\exp\left(\left(\log\left(\pi_i\right) + g_i\right)/\tau\right)}{\sum_{j=1}^{h_{disc}} \exp\left(\left(\log\left(\pi_j\right) + g_j\right)/\tau\right)} \quad \text{for } i = 1, \ldots, h_{disc}. \tag{3}$$

Method



$$\hat{z} = FC_{\hat{z}}([X_c || X_t])$$

$$\hat{x} = LMHead(p_\theta(\hat{z})) \tag{4}$$

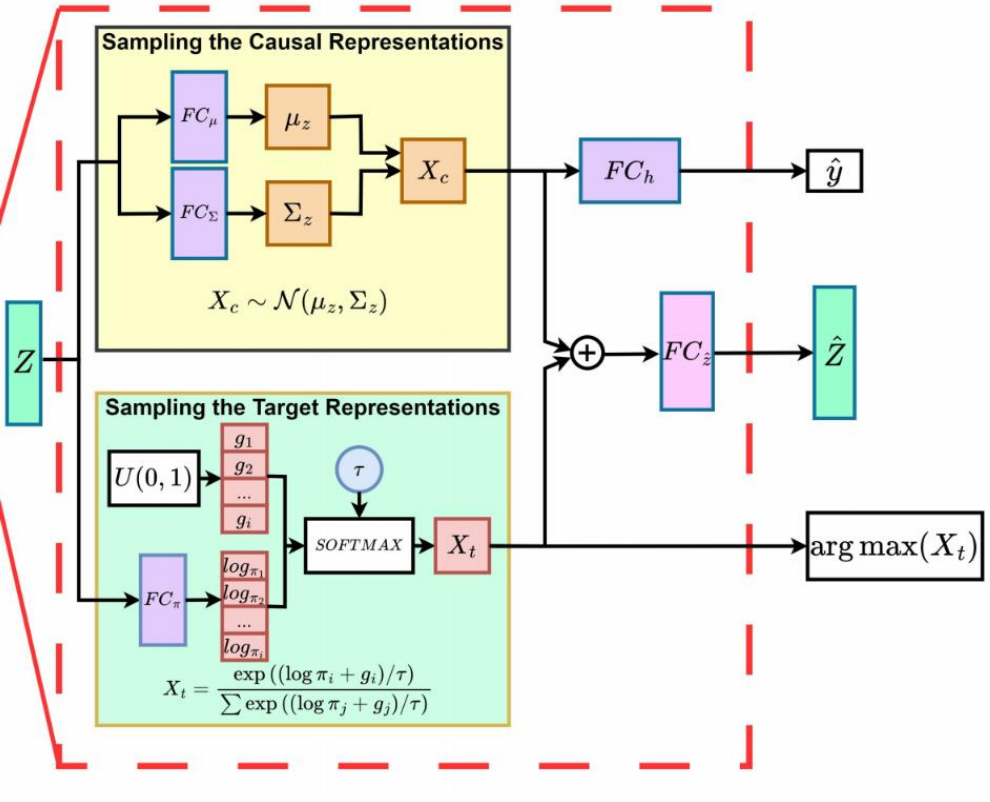
$$\mathcal{L}_{recon}(\gamma(x), \hat{x}) = -\sum_{i=1}^{S} \gamma(x) \log(\hat{x}_i) \tag{5}$$

$$\mathcal{L}_{VAE} = \mathcal{L}_{recon} + \alpha_t * \mathrm{L}_{\mathbb{D}_{target}} + \alpha_c * \mathrm{L}_{\mathbb{D}_{causal}}, \tag{6}$$

$$\mathrm{L}_{\mathbb{D}_{target}} = -D_{\mathrm{KL}}\left(Enc_2(X_t \mid X) || p(X_t)\right) + \alpha_{tc} * \mathcal{L}_{CE}(\arg\max(X_t), t) \tag{7}$$

$$\mathrm{L}_{\mathbb{D}_{causal}} = -D_{\mathrm{KL}}\left(Enc_1(X_c \mid X) || p(X_c)\right) \tag{8}$$

Method

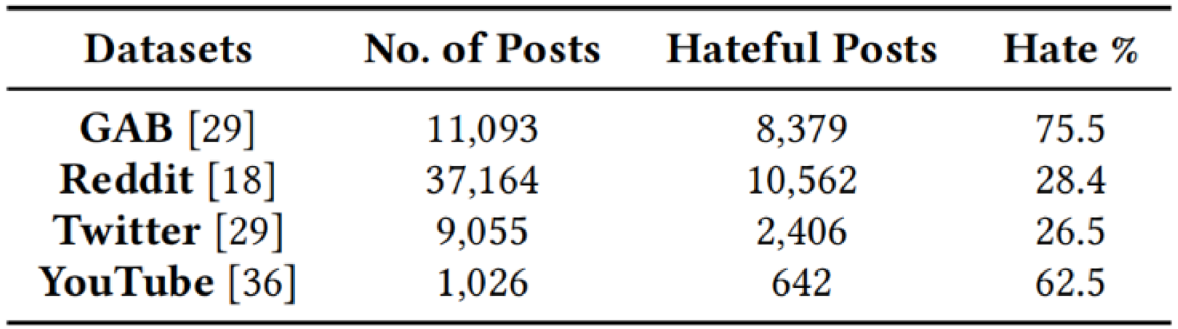


$$\hat{y}_i = \text{Softmax}(FC_h(X_c)) \quad (9)$$

$$\mathcal{L}_{hate} = -\frac{1}{N} \sum_{i=1}^{|D_{source}|} y_i \log \hat{y}_i \quad (10)$$

$$\mathcal{L} = \mathcal{L}_{hate} + \mu_d \mathcal{L}_{VAE} \quad (11)$$

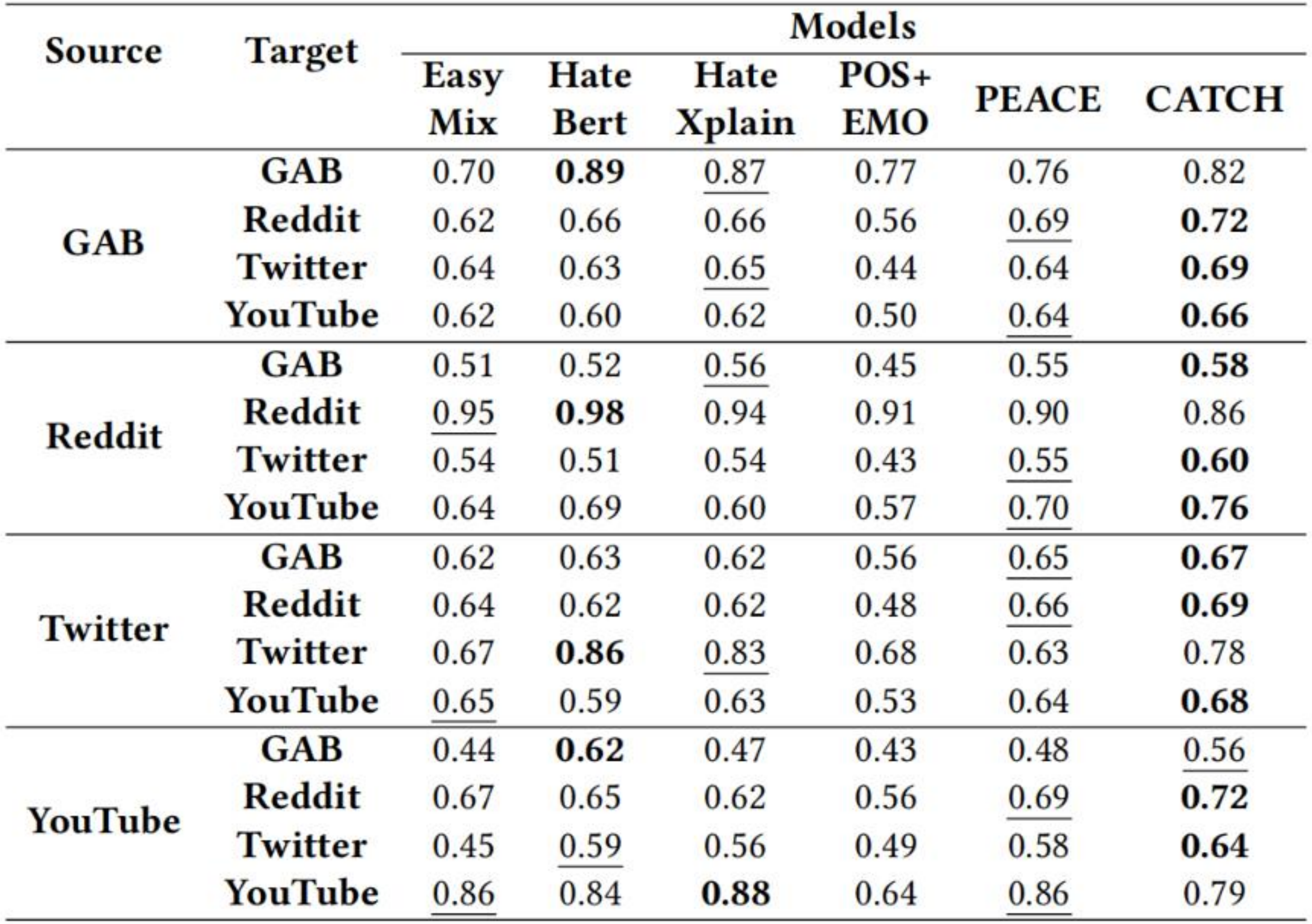
Experiments



Datasets	No. of Posts	Hateful Posts	Hate %
GAB [29]	11,093	8,379	75.5
Reddit [18]	37,164	10,562	28.4
Twitter [29]	9,055	2,406	26.5
YouTube [36]	1,026	642	62.5

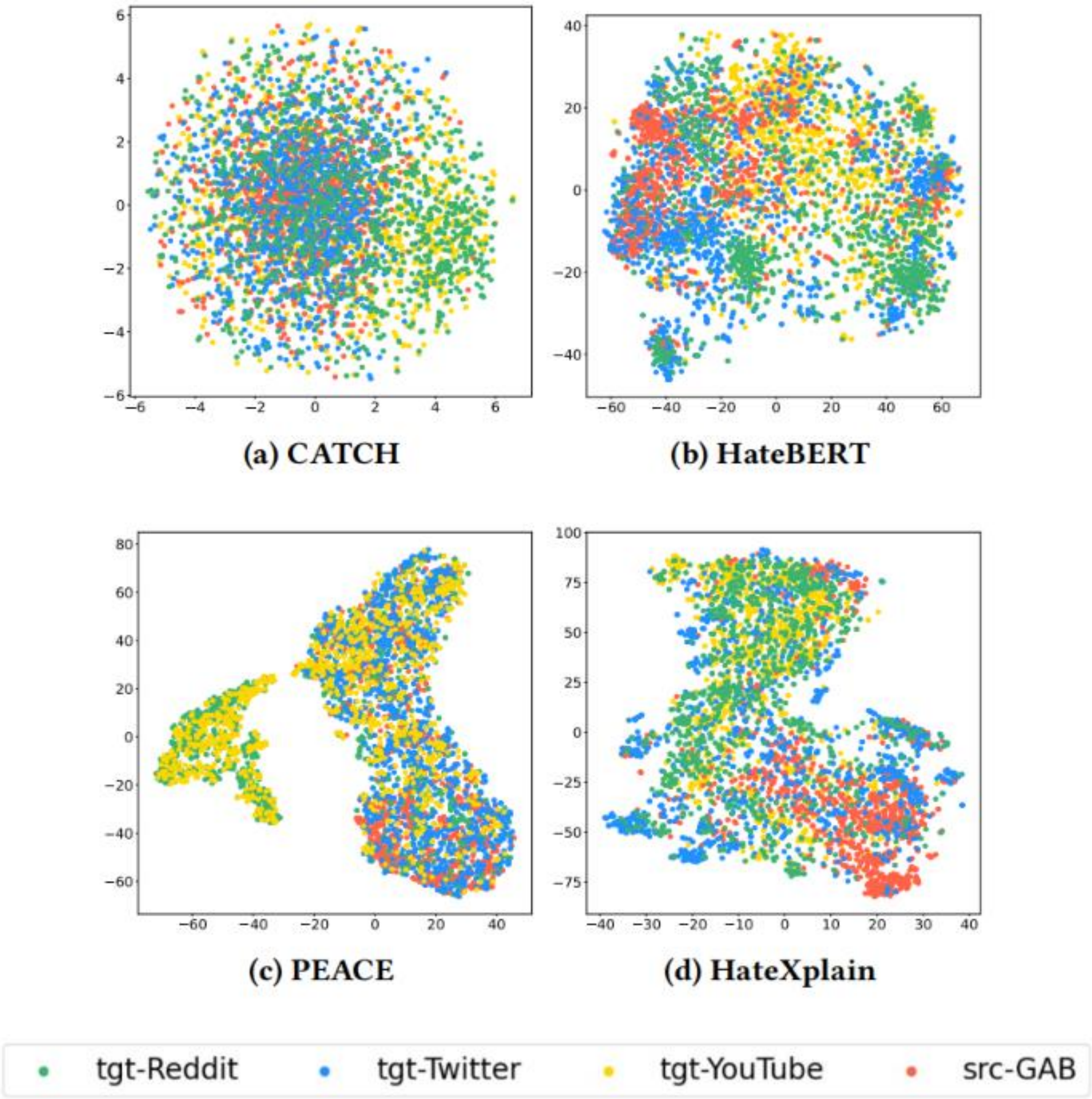
Table 2: Dataset statistics with corresponding platforms and percentage of hateful comments or posts.

Experiments



Source	Target	Models					
		Easy Mix	Hate Bert	Hate Xplain	POS+ EMO	PEACE	CATCH
GAB	GAB	0.70	**0.89**	0.87	0.77	0.76	0.82
	Reddit	0.62	0.66	0.66	0.56	0.69	**0.72**
	Twitter	0.64	0.63	0.65	0.44	0.64	**0.69**
	YouTube	0.62	0.60	0.62	0.50	0.64	**0.66**
Reddit	GAB	0.51	0.52	0.56	0.45	0.55	**0.58**
	Reddit	0.95	**0.98**	0.94	0.91	0.90	0.86
	Twitter	0.54	0.51	0.54	0.43	0.55	**0.60**
	YouTube	0.64	0.69	0.60	0.57	0.70	**0.76**
Twitter	GAB	0.62	0.63	0.62	0.56	0.65	**0.67**
	Reddit	0.64	0.62	0.62	0.48	0.66	**0.69**
	Twitter	0.67	**0.86**	0.83	0.68	0.63	0.78
	YouTube	0.65	0.59	0.63	0.53	0.64	**0.68**
YouTube	GAB	0.44	**0.62**	0.47	0.43	0.48	0.56
	Reddit	0.67	0.65	0.62	0.56	0.69	**0.72**
	Twitter	0.45	0.59	0.56	0.49	0.58	**0.64**
	YouTube	0.86	0.84	**0.88**	0.64	0.86	0.79

Experiments



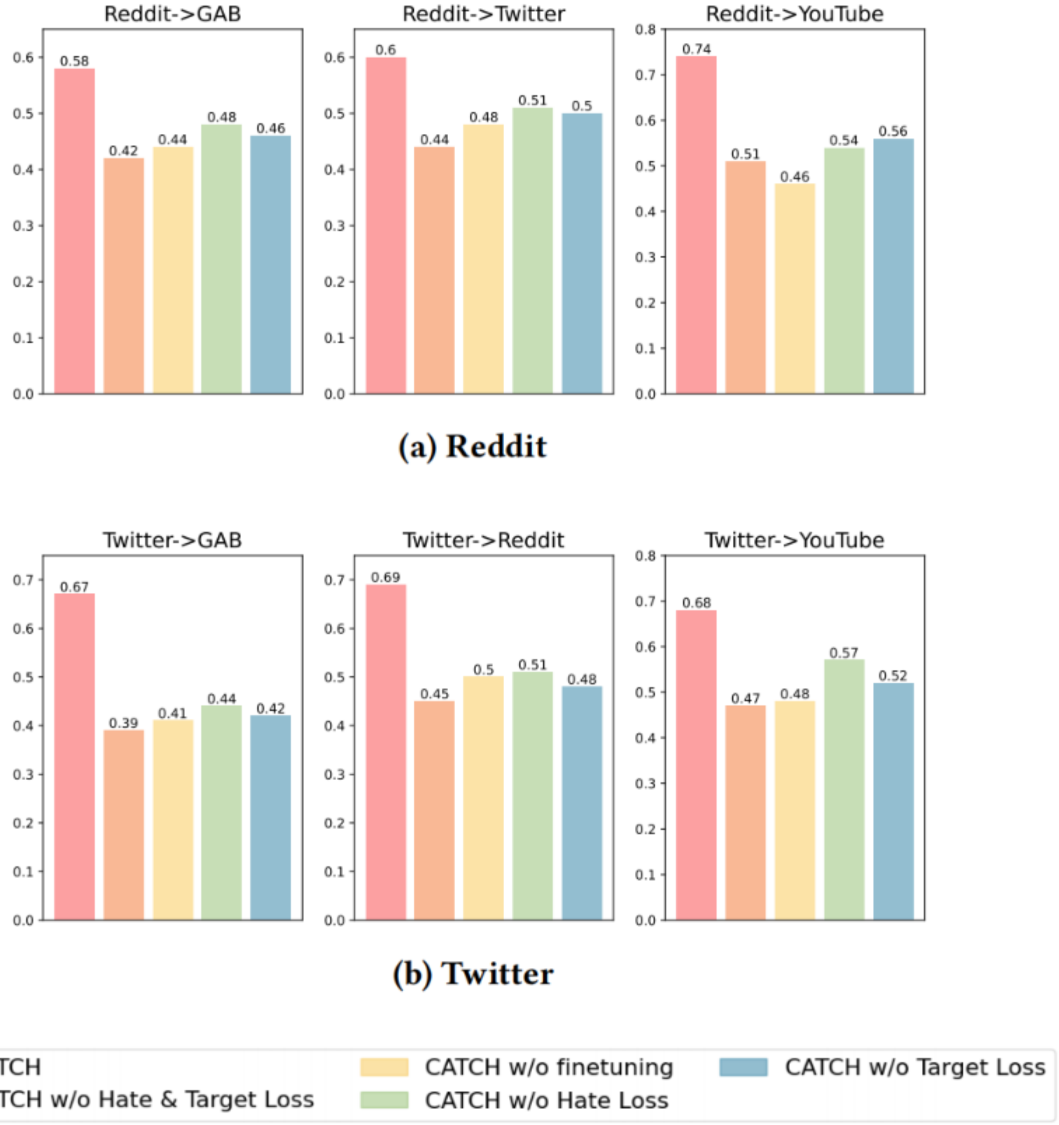
(a) CATCH

(b) HateBERT

(c) PEACE

(d) HateXplain

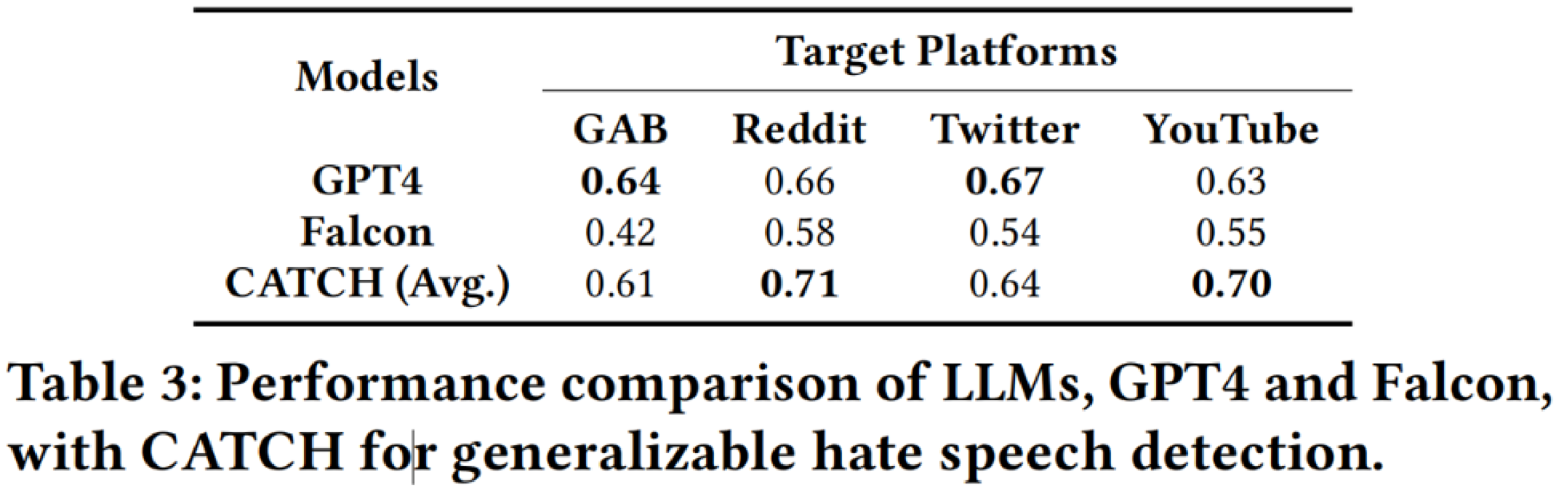
Experiments



(a) Reddit

(b) Twitter

Experiments



Models	Target Platforms			
	GAB	Reddit	Twitter	YouTube
GPT4	**0.64**	0.66	**0.67**	0.63
Falcon	0.42	0.58	0.54	0.55
CATCH (Avg.)	0.61	**0.71**	0.64	**0.70**

Table 3: Performance comparison of LLMs, GPT4 and Falcon, with CATCH for generalizable hate speech detection.

Thanks!